(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 101876963 B
(45)授权公告日 2013.10.16

(21)申请号 201010158624.8
(22)申请日 2010.03.31
(30)优先权数据
087994/2009 2009.03.31 JP
(73)专利权人 株式会社东芝
地址 日本东京都
(72)发明人 村上真之 竹原润 荒牧成彦
川村敏和 高柳洋一 冈部基彦
(74)专利代理机构 永新专利商标代理有限公司 72002
代理人 黄剑锋
(51)Int.Cl.
G06F 13/42(2006.01)
G06F 11/07(2006.01)
(56)对比文件
US 2007217445 A1,2007.09.20, 全文.
CN 1777083 A,2006.05.24, 全文.
CN 101218571 A,2008.07.09, 全文.
审查员 王亮

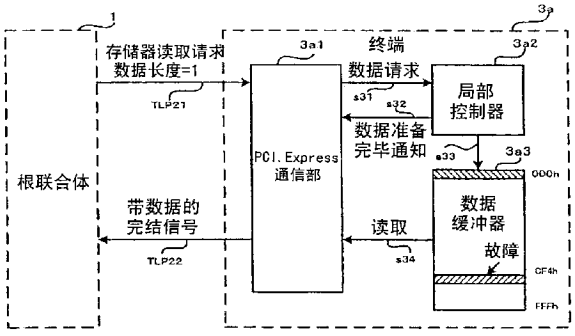
权利要求书2页 说明书10页 附图7页

(54)发明名称

PCI.Express 通信系统及其通信方法

(57)摘要

本发明提供 PCI.Express 通信系统及其通信方法，该方法由以下步骤构成：在 TLP 摘要中，事务层的电路检测错误，对于发送数据设定错误信息，接收到根联合体 (1) 发送的存储器读取请求的终端 (3a) 在对应于被请求的 TLP 的第 1 数据的发送中检测出错误的情况下，将错误信息设置到 TLP 摘要中而返回带数据的完结信号的步骤；根联合体 (1) 将基于错误信息的存储器读取请求发送给终端的步骤；终端返回被请求的第 2 数据的步骤；以及根联合体在保持的第 1 数据的错误位置上覆盖第 2 数据而完成该应答的步骤。



CN 101876963 B

1. 一种 PCI.Express 通信方法，其涉及 PCI.Express 通信系统的发生错误时的传送协议的改良，其特征在于，包括如下步骤：

不对事务层数据包 TLP 的 TLP 摘要附加 ECRC，对于发送数据，通过事务层的电路检测错误，并设定预定格式的错误信息；

发送设备发送存储器读取请求的 TLP；

接收到上述发送设备发送的存储器读取请求的 TLP 的接收设备在与上述发送设备发送的存储器读取请求的 TLP 对应的第 1 数据的发送中检测出错误的情况下，不使返回的完结信号的 TLP 无效化，而在上述 TLP 摘要中，作为上述错误信息而设定错误有无、地址级别单一错误或块级别单一错误、以及错误位置的各位并返回；

上述发送设备根据接收到的上述错误信息，对于发生了错误的地址，对上述地址级别单一错误或上述块级别单一错误，设定预定的各个错误判断级别的最小数据长度，并向上述接收设备请求存储器读取请求；

上述接收设备对上述发送设备向该接收设备请求的存储器读取请求返回被请求的第 2 数据；以及

上述发送设备接收所请求的上述第 2 数据，在所保持的上述第 1 数据的错误位置上覆盖第 2 数据，完成该事务，

通过上述错误信息，能够实现发生错误时的容错性的提高和故障恢复时间缩短。

2. 如权利要求 1 所述的 PCI.Express 通信方法，其特征在于，

设定在上述 TLP 摘要中的上述错误信息包括：

错误检测用冗余码位，上述 TLP 摘要检测本设备的内部电路的故障，设定该 TLP 的数据完全性；

错误有无位，设定在上述 TLP 的数据中是否包含有错误；

错误修复要否位，在上述接收设备发送带数据的完结信号的情况下，对上述发送设备指示是否需要进行该数据的修复，在上述发送设备发送存储器写入请求或消息的情况下，对上述接收设备指示是否要进行错误数据的处理；

错误修复可否位，表示检测出的错误在上述接收设备侧是否能够修复；

地址级别单一错误位，表示上述 TLP 的数据的错误是否只是数据长度 1DW；

块级别单一错误位，表示在上述 TLP 的数据中，是否仅在由预定的地址边界划分的一个块中有数据错误；

错误位置位，表示上述 TLP 的数据的错误的发生位置；以及

奇偶校验位，表示除了上述错误检测用冗余码位以外的上述 TLP 摘要的多个位的奇偶校验。

3. 一种 PCI.Express 通信系统，其涉及 PCI.Express 通信系统的发生错误时的传送协议的改良，其特征在于，包括：

不对事务层数据包 TLP 的 TLP 摘要附加 ECRC，而通过事务层的电路检测错误，并设定预定的发送数据的错误信息的装置；

发送设备，发送存储器读取请求的 TLP；

接收设备，接收上述发送设备发送的存储器读取请求的 TLP，在与上述发送设备发送的存储器读取请求的 TLP 对应的第 1 数据的发送中检测出错误的情况下，不使返回的完结信

号的 TLP 无效化，而在上述 TLP 摘要中，作为上述错误信息而设定错误有无、地址级别单一错误或块级别单一错误、以及错误位置的各位并返回；

上述发送设备还包括根据接收到的上述错误信息，对于发生了错误的地址，对上述地址级别单一错误或上述块级别单一错误，设定预定的各个错误判断级别的最小数据长度，并对上述接收设备请求存储器读取请求的装置；

上述接收设备还包括对上述发送设备向该接收设备请求的存储器读取请求返回被请求的第 2 数据的装置；

上述发送设备还包括接收所请求的上述第 2 数据，在所保持的上述第 1 数据的错误位置上覆盖第 2 数据，完成该事务的装置；

通过上述错误信息，能够实现发生错误时的容错性的提高和故障恢复时间缩短。

4. 如权利要求 3 所述的 PCI.Express 通信系统，其特征在于，

设定在上述 TLP 摘要中的上述错误信息包括：

错误检测用冗余码位，上述 TLP 摘要检测本设备的内部电路的故障，设定该 TLP 的数据完全性；

错误有无位，设定在上述 TLP 的数据中是否包含有错误；

错误修复要否位，在上述接收设备发送带数据的完结信号的情况下，对上述发送设备指示是否需要进行该数据的修复，在上述发送设备发送存储器写入请求或消息的情况下，对上述接收设备指示是否要进行错误数据的处理；

错误修复可否位，表示检测出的错误在上述接收设备侧是否能够修复；

地址级别单一错误位，表示上述 TLP 的数据的错误是否只是数据长度 1DW；

块级别单一错误位，表示在上述 TLP 的数据中，是否仅在由预定的地址边界划分的一个块中有数据错误；

错误位置位，表示上述 TLP 的数据的错误的发生位置；以及

奇偶校验位，表示除了上述错误检测用冗余码位以外的上述 TLP 摘要的多个位的奇偶校验。

PCI.Express 通信系统及其通信方法

[0001] 相关申请的交叉引用

[0002] 本申请基于 2009 年 3 月 31 日提交的日本专利申请第 JP2009-87994 号，要求享受其优先权，并在这里引用其全部内容。

技术领域

[0003] 本发明涉及 PCI.Express 通信系统及其通信方法，特别涉及发生错误时的事务层数据包 (TLP :Transaction Layer Packet) 的传送协议的改良。

背景技术

[0004] PCI.Express（注册商标）总线是为了传送计算机系统及其他电子设备的数据而近年来开发的基于点对点连接的高速串行接口，与以往的并行传送相比总线的基板占用面积较少而能够实现进一步的小型化，在许多领域中研究了其用途。

[0005] 其规格的详细情况由作为 PCI 规格的制定基础的 PCI-SIG(PeripheralComponent Interconnect-Special Interest Group :互连外围设备专业组）规格化为 PCI.Express Base Specification。此外，还出版了该规格的说明书。例如有日本的技术书籍《PCIe 入门讲座》（荒井信隆，里美尚志，田中显裕共著，（株）电波新闻社，2007 年 4 月 1 日发行，第 1 ～第 5 章）（以下称作非专利文献 1)。

[0006] 首先，参照图 1 至图 3，说明该 PCI.Express 通信系统（有时也称作 PCIe 通信系统）的概况。PCI.Express 通信系统结构例如如图 1 所示，由根联合体 (Root Complex)1、开关 2 及终端 3(3a、3b、3c、3d) 的设备构成。

[0007] 此外，根联合体 1、开关 2 具有多个端口，将它们与终端 3 的相互间连接的 PCI.Express 总线 7a ～ 7e 具备如图 2 所示的 3 层的层构造。

[0008] 各个层包括 :事务层 101，对由最上位的驱动程序及应用软件构成的上位软件层，用以往的 PCI 互换服务和端对端来保证数据的可靠的通信 ;数据链路层 102，保障相邻的组件间的可靠的数据通信 ;以及物理层 103，在物理介质上收发通信数据包，并且上述各个层将收发的数据以数据包的形式传送。

[0009] 进而，根联合体 1 位于 PCI.Express 通信系统的树构造的最上位，经由系统总线（未图示标号）与 CPU5 连接，并经由存储器总线（未图示标号）与存储器 6 连接。

[0010] 在该结构中，在根联合体 1 与终端 3a 的通信中，开关 2 为 TLP 的中继设备，在终端 3a 与终端 3d 的通信中，开关 2 和根联合体 1 为中继设备。

[0011] 这样构成的 PCI.Express 通信系统的设备间的传送路径的连接为点对点连接，使用两个单向的差动放大器的双重单工方式，链路速度具有 2.5Gbps 的带宽，双向具备 5Gbps 的频带。

[0012] 进而，通过将该双向的传送路径（称作通道 (lane)）从两组增加到 32 组，能够将总线的频带宽度可扩展地构成，通过在该传送路径上收发数据包来执行数据的传送。

[0013] 各层的数据包如图 1 所示，在事务层及数据链路层中生成，分别称作事务层数

据包 (TLP，Transaction Layer Packet)、数据链路层数据包 (DLLP，Data Link Layer Packet)。

[0014] 此外，在物理层中也为了链路控制而生成物理层数据包 (PLP，PhysicalLayer Packet)。

[0015] 此外，各层的数据包在由链路连接的对方的相同层之间交换，如图 3 所示，在下位的协议层中在前后附加了信息，最终发送给传送路径。接收到的数据包在各协议层中删除前后的信息，传递给上位的协议层。

[0016] 详细地讲，进行端对端通信的 TLP 在事务层中构成 TLP 标头、数据有效载荷、以及可选的 TLP 摘要（称作 ECRC，End-to-end CRC，端对端循环冗余校验），在数据链路层中，在发送时附加顺序号码和 LCRC（链路循环冗余校验码），在接收时检查后删除。

[0017] DLLP 是 TLP 的送达应答（肯定应答 Ack 和否定应答 Nak）等在链路的双方间交换信息的短数据包。

[0018] 进而，在各 TLP 中，对数据包的两端附加用来在物理层的接收侧检测 TLP 开始和结束的控制字符 (STP 和 END)。此外，在各 DLLP 中，对数据包的两端附加用来检测 DLLP 开始和结束的控制字符 (SDP 和 END)。

[0019] 接着，对于这样构成的 PCI.Express 通信系统的发生错误时的事务层中的错误处理的问题，参照图 4A ～图 4D 及图 5 进行说明。

[0020] 图 4A ～图 4D 是说明在由根联合体 1、开关 2 及终端 3a ～ 3c 构成的 PCI.Express 通信系统中，在从终端 3a 发送发送数据中发生了错误的情况下的问题的图，此外，图 5 是说明发生错误时的容错功能的问题的图。

[0021] 例如，PCIe 设备的结构如图 5 所示的终端 3，包括 :PCI.Express 通信部 3a1 ;局部控制器 3a2，从 PCI.Express 通信部 3a1 接收发送数据的请求，控制向对通信的发送数据进行存储的数据缓冲器（存储器）3a3 发送的数据的写入 ;以及数据缓冲器 3a3。

[0022] 由于从 PCI.Express 设备发送的发送数据是健全的，所以通常需要对于从数据缓冲器 3a3 读出到 PCI.Express 通信部 3a1 中的数据、在 PCI.Express 通信部 3a1 侧检测错误。在该错误检测中，检测局部控制器 3a2、数据缓冲器 3a3、以及 PCI.Express 通信部 3a1 与局部控制器 3a2 及数据缓冲器 3a3 之间的接口等的 PCI.Express 通信部 3a1 的上位电路的硬件故障、还有起因于数据缓冲器 3a3 的软件错误等的发送数据的错误。

[0023] 首先，对事务层的数据传送进行叙述。在 TLP 的标头中准备了 EP 位。对于 TLP 的发送数据，例如用作为如上述的 PCI.Express 通信部 3a1 的事务层的电路的一部分设置的错误检测电路来检测错误，在该错误是不能修复的情况下，事务层的错误检测电路在 EP 位中设置 1 而发送 TLP，则接收侧通过参照 EP 位，了解在接收数据中包含有错误，所以能够保障端对端的数据完全性。

[0024] 但是，发送侧的事务层的错误检测电路为了在标头的 EP 位中设置 1，需要先将发送数据全部储存到数据缓冲器中，使 PCI.Express 通信系统的吞吐量下降。

[0025] 如果上述 PCI.Express 的上位电路的总线的频带比 PCI.Express 的频带高，则优选地，事务层的电路将从上位电路传送来的发送数据不先储存到数据缓冲器中而从 PCI.Express 的传送路径依次传送。将该传送方式称作直通 (cut through)。

[0026] 接着，参照图 4A ～图 4D，说明采用了该直通传送方式的设备中的错误处理的延

迟成为问题的情况。图 4A 是图示了在从终端 3a 对根联合体 1 发送带数据的完结信号(completion)TLP1 过程中、终端 3b 对根联合体 1 开始发送 1024 双字（以后记作 DW）的 TLP2、在终端 3a 内检测出了在 TLP1 的未发送的数据中有不能修复的错误的情况下的状态的图。

[0027] 在此情况下，终端 3a 如图 4B 所示，在 TLP 的末尾的控制字符中不是附加“END”而是附加“EDB”(EnD Bad) 使 TLP1 无效化，接着要对根联合体 1 发送表示该错误是致命性的的错误消息的 TLP3。

[0028] 从终端 3b 发送的 TLP2 在 TLP1 的发送结束为止在开关 2 内的缓冲器内处于待机状态，所以如图 4C 所示，如果 TLP1 的发送结束，则开关 2 对根联合体 1 发送 TLP2。由于 TLP2 是发送中，所以错误消息的 TLP3 在开关 2 的缓冲器内待机。

[0029] 接着，如图 4D 所示，如果 TLP2 的发送完毕，则开始对根联合体 1 发送 TLP3。

[0030] 在上述那样的情况中，错误消息的 TLP3 在 2.5Gbps、通道数 1 的情况下，有发生最大约 16μsec(1024DW×16ns/DW) 的延迟的问题。当为开关 2 将许多终端连接的系统的情况下，该延迟有可能变得更大。

[0031] 此外，在终端 3c 构成为终端 3a 的待机类设备的情况下，在将该错误消息的 TLP3 触发而执行设备的切换的情况中，切换时间的延迟作为系统的故障时间而成为问题。

[0032] 接着，参照图 5，对发生错误时的系统的容错功能的下降成为问题的情况进行说明。

[0033] 图 5 是说明在根联合体 1 对具备局部控制器 3a2 的终端 3a 发送 1024DW 的存储器读取请求的 TLP31、且终端 3a 发送对该请求的完结信号的 TLP32 的过程中、在数据缓冲器 3a3 内的未发送数据中检测出错误的情况下的终端 3a 的错误修复动作的结构图。

[0034] 在 PCI.Express 设备中，在事务层的数据传送中使用直通的情况下，通常为下述这样的修复动作。

[0035] (1) 终端 3a 的 PCI.Express 通信部 3a1 接收存储器读取请求的 TLP31，对局部控制器 3a2 请求 1024DW 的数据 (s41)。

[0036] (2) 局部控制器 3a2 将数据传递给数据缓冲器 3a3(s43)，并对 PCI.Express 通信部 3a1 通知数据准备完毕 (s42)。

[0037] (3)PCI.Express 通信部 3a1 一边从数据缓冲器 3a3 进行突发读取 (burstread)(s44)，一边发送完结信号 TLP32。

[0038] 此时，在数据缓冲器地址 CF4h 的数据中检测出错误，附加完结信号的 TLP32 的末尾字符“EDB”，使发送中的 TLP32 无效化。

[0039] (4) 并且，为了进行数据缓冲器 3a3 的错误修复，PCI.Express 通信部 3a1 对局部控制器 3a2 再次请求数据。

[0040] 这里，对于终端 3a 的数据缓冲器 3a3 的错误的修复动作，分为错误处理对应为以一时性的软件错误为前提设计的系统的情况、和以不能修复的永久故障为前提设计的系统的情况说明其问题。

[0041] 在前者的情况下，通常，PCI.Express 通信部 3a1 同样对局部控制器 3a2 请求 1024DW 的数据。在此情况下，在缓冲器地址 CF4h 是永久故障的情况下，由于相同的数据被写入到数据缓冲器 3a3 的的相同区域中，所以有地址 CF4h 的数据的错误不能被修复的问

题。

[0042] 此外，由于没有将正常的读取数据返回给根联合体 1，所以 PCI.Express 通信部 3a1 将不能恢复的错误 (Fatal Error：致命错误）的错误消息发送给根联合体 1。

[0043] 并且，根联合体 1 通过接收该错误消息，将系统停止或复位。

[0044] 在后者的情况下，有 PCI.Express 通信部 3a1 对局部控制器 3a2 以较小的数据尺寸请求数据的发送的方法，将完结信号 TLP 分多次发送给根联合体 1。这有事务的完成需要较长时间的问题。

[0045] 在除此以外的方法中，还有将数据缓冲器 3a3 双路化而交替地使用的方法，但耗费存储器的成本。

[0046] 关于该 PCI.Express 的 TLP，公开了将数据的错误检测委托给数据链路层的功能的 LCRC、不使用事务层的功能的 ECRC 而独自利用 TLP 摘要的字段的技术。例如，美国专利申请公开第 2009/0006932 号说明书（以下称作专利文献 1）。

[0047] 在 PCI.Express 规格中，作为用来保证端对端的数据完全性的可选规格，将 ECRC 保存到 TLP 摘要中。

[0048] 但是，如在专利文献 1 中所记载，如果充分确保中继设备的可靠性，再在事务层中设置数据错误检测功能，例如对收发的数据缓冲器设置奇偶校验 (parity)，则能够通过 LCRC 补充端对端的数据完全性的保证，可以判断为不需要 ECRC。

[0049] 专利文献 1 的作为用来独自利用 TLP 摘要的实施例，用 TLP 的标头部的 TD 位表示 TLP 摘要的存在的有无，利用标头的保留位表示 TLP 摘要是保存着独自的信息、还是保存着 ECRC。

[0050] 但是，在以这样的目的利用保留位而构建 PCI.Express 通信系统的情况下，在将来 PCI.Express 被修订、对其保留位分配了新的定义的时候，有可能不能取得与将来的 PCI.Express 规格的互换性。

[0051] 因而，在独自使用 TLP 摘要的情况下，需要能够不使用标头的保留位、而在系统内的 PCIe 设备中共享 TLP 摘要是独自规格的情况。

[0052] 如以上说明，关于基于以往的 PCI.Express 规格的 TLP 事务，在错误发生时的故障恢复时间、以及容错性 (fault tolerance) 方面存在问题。

发明内容

[0053] 本发明是为了解决上述问题而做出的，目的是提供一种是 PCI.Express 通信系统中的发生错误时的 TLP 传送协议的改良、对于 PCI.Express 通信系统的中继设备或终端设备内的数据缓冲器内的特定（规定）区域的不能恢复的故障、一时性的软件错误、能够实现 PCI.Express 通信系统的容错性的提高、并且能够缩短故障恢复时间的 PCI.Express 通信系统及其通信方法。

[0054] 为了达到上述目的，本发明的 PCI.Express 通信系统的通信方法由以下的结构构成。即，

[0055] 其是 PCI.Express 通信系统的发生错误时的传送协议的改良，其特征在于，包括：

[0056] 不对事务层数据包 (TLP) 的 TLP 摘要附加 ECRC，对于发送数据，通过事务层的电路检测错误，并设定预定格式的错误信息的步骤；

[0057] 发送设备（请求方）发送存储器读取请求的 TLP；

[0058] 接收到上述 TLP 的接收设备（完结方）在与被请求的上述 TLP 对应的第 1 数据的发送中检测出错误的情况下，不使返回的完结信号 TLP 无效化，而在上述 TLP 摘要中，作为上述错误信息而设定错误有无、地址级别单一错误或块级别单一错误、以及错误位置的各位并返回的步骤；

[0059] 上述发送设备根据接收到的上述错误信息，对于发生了错误的地址，对上述地址级别单一错误或上述块级别单一错误，设定预定的各个错误判断级别的最小数据长度，并向上述接收设备请求存储器读取请求的步骤；

[0060] 上述接收设备对上述存储器读取请求返回被请求的第 2 数据的步骤；以及

[0061] 上述发送设备接收所请求的上述数据，在所保持的上述第 1 数据的错误位置上覆盖第 2 数据，完成该事务，

[0062] 通过上述错误信息，能够实现发生错误时的容错性的提高和故障恢复时间缩短。

[0063] 进而，为了达到上述目的，本发明的 PCI.Express 通信系统由以下的结构构成。即，

[0064] 是将 PCI.Express 通信系统的发生错误时的传送协议改良的系统，其特征在于，

[0065] 构成上述 PCI.Express 通信系统的收发 TLP 数据包的设备不对事务层数据包 (TLP) 的 TLP 摘要附加 ECRC，而通过事务层的电路检测错误，并设定预定的发送数据的错误信息；

[0066] 发送设备（请求方）发送存储器读取请求的 TLP；

[0067] 接收到上述 TLP 的接收设备（完结方）在与被请求的上述 TLP 对应的第 1 数据的发送中检测出错误的情况下，不使返回的完结信号的 TLP 无效化，而在上述 TLP 摘要中，作为上述错误信息而设定错误有无、地址级别单一错误或块级别单一错误、以及错误位置的各位并返回；

[0068] 上述发送设备根据接收到的上述错误信息，对于发生了错误的地址，对上述地址级别单一错误或上述块级别单一错误，设定预定的各个错误判断级别的最小数据长度，并对上述接收设备请求存储器读取请求；

[0069] 上述接收设备对上述存储器读取请求返回被请求的第 2 数据；

[0070] 上述发送设备接收所请求的数据，在保持的上述第 1 数据的错误位置上覆盖第 2 数据，完成该事务；

[0071] 通过上述错误信息，能够实现发生错误时的容错性的提高和故障恢复时间缩短。

[0072] 根据本发明，能够提供一种对于 PCI.Express 通信系统的中继设备或终端设备内的数据缓冲器内的特定区域的不能恢复的故障及一时性的软件错误、能够实现 PCI.Express 通信系统的容错性的提高、并且能够缩短故障恢复时间的 PCI.Express 通信系统及其通信方法。

附图说明

[0073] 图 1 是以往的 PCI.Express 通信系统的结构图。

[0074] 图 2 是说明以往的 PCI.Express 的构造的图。

[0075] 图 3 是说明以往的 TLP 格式的图。

[0076] 图 4A ～图 4D 是说明以往的 PCI.Express 通信系统的问题的图。

[0077] 图 5 是说明以往的 PCI.Express 通信系统的问题的图。

[0078] 图 6 是本发明的 PCI.Express 通信系统的 TLP 的格式的结构图。

[0079] 图 7 是说明本发明的 PCI.Express 通信系统的图。

[0080] 图 8 是说明本发明的基于 TLP 摘要的通信动作的流程图。

[0081] 图 9 是说明本发明的基于 TLP 摘要的通信动作的流程图。

[0082] 图 10 是本发明的使用 TLP 摘要的 PCI.Express 通信系统的实施例。

具体实施方式

[0083] 以下，参照附图对本发明的实施例进行说明。

[0084] [实施例 1]

[0085] 在图 6 中表示本发明的 TLP 摘要的格式的例子，并且，在图 7 中表示使用该 TLP 摘要的 PCI.Express 通信系统的动作例的结构图。

[0086] 对该实施例 1 的各部，将与图 1 ～图 5 所示的以往的实施例的 PCI.Express 通信系统相同的部分用相同的符号表示，并省略其说明。

[0087] 图 6 的本发明的实施例与在图 5 中说明的以往的 PCI.Express 通信系统的实施例的不同点是，在采用直通传送方式的以往的 PCI.Express 通信系统中，在终端 3a 发送 TLP 数据包的过程中在未发送的数据中检测出了错误的情况下，一般的方法是使 ECRC(端对端循环冗余校验) 反转、将 TLP 的末尾的控制字符设为“EDB”而不是“END”、使数据包无效化 (NullifiedTLP)，但在本发明中，将预先设定的错误信息保存到 TLP 摘要中，对于在作为 PCI.Express 通信部 3a1 的上位电路的局部控制器 3a2、数据缓冲器 3a3，以及这些电路与 PCI.Express 通信部 3a1 的接口中发生的一时性的错误、在事务层的数据缓冲器 3a3 中发生的软件错误、以及特定位置的不能恢复的故障，实现 PCI.Express 通信系统的容错性的提高，并且缩短故障恢复时间。

[0088] 首先，参照图 6，说明本申请的事务层数据包的构造和对数据包末尾附加的 32 位的 TLP 摘要的格式的设定例。

[0089] 设定在位 16 ～ 31 的高位 16 位的错误检测冗余码 D1 是设置用来保证端对端的数据完全性的冗余码的区域。在端对端的中继设备的可靠性不详、或不充分的情况下，为了保证端对端的数据完全性而设置该区域。它相当于 PCI.Express 的 ECRC 的作用，除了 16 位的 CRC 以外，也可以设定 16 位 SUM 等。

[0090] 对位 15 设定的奇偶校验 (parity)D2 是用来保证后述的位 0 ～ 14 的数据的奇偶校验位。虽然能够通过 LCRC(链路循环冗余校验码) 检测数据的错误，但 TLP 摘要的字段是用来进一步提高数据的可靠性的。

[0091] 对位 14 设定的错误有无 D3 是表示在发送的 TLP 的数据中是否包含有错误 (1 :有错误，0 :无错误) 的位。

[0092] 此外，对位 13 设定的错误修复要否 D3 在接收到带数据的完结信号 (completion) 的情况下，指示在请求侧是否需要错误数据的修复处理。

[0093] (例如，1 :请求方 (发送设备) 仅对包含错误数据的地址、或其周边地址区域发出最小的数据长度的存储器读取请求、或者发出与上次相同的地址和数据长度的存储器读取

请求。

[0094] 0:请求方不论错误的有无都将读取的数据进行通常处理,结束事务。)

[0095] 在接收到存储器写入请求或消息的情况下,是对完结方(接收设备)侧指示错误数据的处理的位。

[0096] (例如,1:完结方将错误数据丢弃。

[0097] 其中,由系统决定是将所有的数据丢弃,还是仅将错误数据丢弃而保持剩余的数据。

[0098] 0:完结方将该写入数据进行通常处理。)

[0099] 接着,对位 12 设定的错误修复可否 D5 是表示在发送侧发生的错误是否能够修复的位。

[0100] (例如,1:通过发送侧的功能能够再生成正确的数据,能够修复。

[0101] 0:通过发送侧的功能不能再现数据,不能修复。)

[0102] 此外,对位 11 设定的地址级别 (address level) 单一错误 D6 是表示在 TLP 的数据中是否包含有只有 1DW 的数据错误的位。

[0103] (例如,1:单一错误,0:多个错误)

[0104] 同样,对位 10 设定的块级别 (block level) 单一错误 D7 是表示在 TLP 的数据中是否仅在由预先在系统设计时设定的地址边界划分的一个块中包含有数据错误的位。

[0105] (例如,1:单一错误,0:多个错误)

[0106] 进而,对位 0 ～ 9 设定的错误位置 D8 是设置最初检测出错误的位置的区域,是表示关于 TLP 的数据有效载荷、从开头起第几个 DW 的错误的位。

[0107] 这里,将关于在图 6 所示的 TLP 摘要的格式中说明的错误的设定信息在这里统称作错误信息。

[0108] 接着,参照图 7 说明基于这样的 TLP 摘要的格式设定的 PCIe 通信系统的事务动作,。

[0109] 图 7 是根据图 5、根联合体 1 对具备局部控制器 3a2 的终端 3a 发出 1024DW 的存储器读取请求 (TLP)、在终端 3a 发送带数据的完结信号的过程中、在数据缓冲器 3a3 内的未发送数据中检测出错误的情况下的终端 3a 的错误修复的情况下的动作的结构图。

[0110] 在本发明中,为以下这样的错误修复动作。首先,与图 5 同样,

[0111] (1) 终端 3a 的 PCI.Express 通信部 3a1 接收存储器读取请求的 TLP31,对局部控制器 3a2 请求 1024DW 的数据。

[0112] (2) 局部控制器 3a2 将数据传递给数据缓冲器 3a3,对 PCI.Express 通信部 3a1 通知数据准备完毕。

[0113] (3)PCI.Express 通信部 3a1 一边从数据缓冲器 3a3 进行突发读取 (s44),一边发送带数据的完结信号的 TLP32。

[0114] 此时,在数据缓冲器地址 CF4h 的数据中检测出错误,在完结信号 TLP 的末尾不附加“EDB”,不使发送中的 TLP 无效化。

[0115] 接着,在终端 3a 的数据缓冲器 3a3 的错误的修复动作中,

[0116] (4) 将从发送中的数据检测出的错误信息以图 6 中说明的 TLP 摘要的格式设定,设定的错误信息例如是单一错误,将记述有其错误位置的地址的 TLP(带数据的完结信号的

TLP32) 从终端发送。

[0117] (5) 由图 7，根联合体 1 在保持着接收到的数据的状态下仅对发生了错误的位置的地址发出存储器读取请求。由于此时的 TLP 的数据长是单一错误，所以在 TLP 标头的数据长字段中设定 1(TLP21)。

[0118] (6) 终端 3a 的 PCI.Express 通信部 3a1 对于局部控制器 3a2，对请求的地址，请求 1DW 的数据 (s31)。

[0119] (7) 根联合体 1 在数据缓冲器 3a3 的开头准备 1DW 的数据 (s33)，对 PCI.Express 通信部 3a1 通知数据准备完毕 (s32)。

[0120] (8) 并且，PCI.Express 通信部 3a1 读取数据缓冲器 3a3 的开头地址的数据 (s34)，将该数据的 TLP(TLP22) 发送给根联合体 1。

[0121] (9) 于是，根联合体 1 接收该 TLP，将该数据覆盖在所保持的数据的错误位置上，结束该事务。

[0122] 在该错误处理的情况下，如果写入被请求的 1DW 的数据的数据缓冲器 3a3 的开头地址故障，则不能发送正确的数据，而在此情况下，通过对根联合体 1 指定以使写入到预先设定的别的区域中而能够避免。

[0123] 在基于设定在该 TLP 摘要中的错误信息进行的错误处理的情况下，不仅是对于一时性的软件错误，对于特定（规定）区域的永久故障也能够实现容错。

[0124] 进而，由于仅对错误位置的特定区域发出存储器读取请求而进行错误修复，所以能够迅速地结束事务。

[0125] 因而，即使是在数据缓冲器 3a3 内发生了单一错误的情况、或在数据缓冲器与局部控制器之间发生了一时性的错误的情况，也有可能能够将错误的修复迅速地修复，所以能够确保 PCI.Express 通信系统对于 PCI.Express 的上位电路的错误的实时性。

[0126] 接着，参照图 8 及图 9，说明本申请的基于设定了错误信息的 TLP 摘要的 TLP 接收处理动作的详细情况。

[0127] 参照奇偶校验 D2 及错误检测用冗余码 D1，判断在接收到的 TLP 中是否没有错误 (s1 ～ s3)。

[0128] 在任一个中有错误的情况下，将该数据包丢弃 (s4)，在该 TLP 的种类是请求的情况下，通过预先设定的系统的恢复方法进行处理 (s5、s6)。

[0129] 否则，即为完结信号的情况下，再发出前次的请求，或者将发送侧设备或系统复位 (s7)。

[0130] 在接收到的 TLP 中没有错误、其种类是带数据的完结信号以外的情况下 (s11)，在其种类是存储器写入请求或消息的情况下 (s12)，再参照该 TLP 的有错误 D3、要错误修复 D4，执行预先设定的错误修复处理 (s15)，否则（没有错误 D3，不要错误修复 D4)，结束该应答。

[0131] 在该 TLP 的种类不是带数据的完结信号、存储器写入请求、消息请求的任一种的情况下，在该 TLP 是无数据完结信号的情况下，结束该应答 (s16)，否则，即如果是存储器写入请求及消息请求以外的请求，则执行完结 (completion)(s17)。

[0132] 接着，参照图 9 说明在该 TLP 是带数据的完结信号的情况下，接收到它的设备中的处理动作。

[0133] 在被请求的数据中没有错误 D3(s21)、错误修复要否 D4(s22) 是不需要的情况下，结束读取事务 (s23)。

[0134] 在数据中有错误 D3(s21)、需要错误修复 D4(s22) 的情况下，再判断错误修复可否 D5(s24)。在错误修复可否 D5 是不能的情况下，参照地址级别单一错误 D6，在其结果是单一错误 (s25) 的情况下，向错误位置 D8 的地址发送数据长 1DW 的存储器读取请求 (s26)。

[0135] 如果正常地返回了对于该请求的完结信号，则结束该读取事务 (s27、s23)。

[0136] 参照地址级别单一错误 D6，在其结果是块级别单一错误的情况下 (s28)，对基于该错误位置 D8 的 nDW(DW 的 n 倍）地址边界的开头地址，发送数据长 nDW 的存储器读取请求 (s29)。如果正常地返回了对于该请求的完结信号，则结束该读取事务 (s27、s23)。

[0137] 在错误修复可否 D5 是否的情况下，或者在既不是地址级别单一错误也不是块级别单一错误的情况下 (s28 :否)，再发出与上次相同的请求，或者将发送侧 PCI.Express 设备或系统整体复位。

[0138] [实施例 2]

[0139] 接着，说明将独自使用如实施例 1 中说明的 TLP 摘要的 PCI.Express 设备（以下称作专用设备）和以往的 PCI.Express 设备（以下称作通用设备）混合的 PCI.Express 通信系统的通信方法。

[0140] 如果独自地设计 PCI.Express 通信系统内的所有的 PCI.Express 设备，则各设备只要设计为使其能够对 TLP 摘要进行如实施例 1 中说明的错误信息的设定、检测就可以。

[0141] 但是，即使是要求可靠性的面向产业的 PCI.Express 通信系统，也可以想到所采用的 PCI.Express 通信系统的结构中混合有通用设备和如实施例 1 所示的独自设计的专用设备的情况。

[0142] 例如，如图 10 所示，根联合体 1x 是用于 CPU5 及存储器 6 的接口的高可靠化、或 CPU5 的双路化控制等的专用设备，终端 3cx、3dx 例如是实时以太网（注册商标）等的专用设备，事务层能够独自地设计。

[0143] 另一方面，终端 3a、3b 例如是通用以太网（注册商标）、或通用 USB，由于它们是 COTS(商用品，Commercial off-the-shelf)，所以不能对事务层的电路加以改动。

[0144] 该专用设备的终端 3cx、3dx 例如在需要实时以太网（注册商标）等的控制通信的情况下，可以独自地设计事务层，但在通用以太网（注册商标）3b 或通用 USB3a 端口的情况下，如果将该设备做成专用设备则变得昂贵，所以对于这些通用设备的 PCI.Express 通信部不加以改动。

[0145] 在上述那样的将通用和专用的设备混合的情况下，根联合体 1x 需要容易地识别哪个 PCI.Express 设备支持如实施例 1 所示的独自的 TLP 摘要。

[0146] 在专利文献 1 中，利用 TLP 标头的保留位，对所发送的每一个 TLP 判别 TLP 摘要的类型是 ECRC 还是独自，但在本实施例 2 中，当完结 (completion) 时一起进行专用设备的判断这一点不同。

[0147] 以下，参照图 10，对混合通用、专用设备的系统结构中的设备的识别方法进行说明。

[0148] 图 10 的根联合体 1x 在系统的初始化时将销售商 (vendor) 独自消息以广播发送。销售商独自消息由 PCI.Express 规格设定，消息码是 0111 1111(Vendor_defined Type

1)。消息的头的第 4DW 是销售商记述字段，预先设定并保存表示独自使用 TLP 摘要的代码。

[0149] 专用设备的终端 3cx 和 3dx 如果接受到该消息，则对根联合体 1x 返回上述销售商独自消息。

[0150] 通用设备的终端 3a 和 3b 由于不能解释销售商独自消息的内容，所以将该消息丢弃。

[0151] 根联合体 1x 将有应答的终端的 ID（总线号码、设备号码、功能号码）全部存储，在初始化结束后，对于与这些 ID 一致的请求及完结信号，独自处理 TLP 摘要，如果不与该 ID 一致，则不使用 TLP 摘要，或者以 ECRC 进行处理。

[0152] 对于广播型的消息请求，专用设备将 TLP 摘要以 ECRC 进行处理、或将该消息的头部的 TD 位设为 0，使 TLP 摘要无效。

[0153] 通过以上说明的方法，能够容易地识别在系统的初始化时构成的设备是通用还是专用的。

[0154] 本发明并不限于如上述实施例的任何限定，对 TLP 摘要设置的错误信息也可以根据缓冲器存储器的结构或数据结构等而变更请求再发送的区域，在不脱离本发明的主旨的范围内能够进行各种变形来实施。

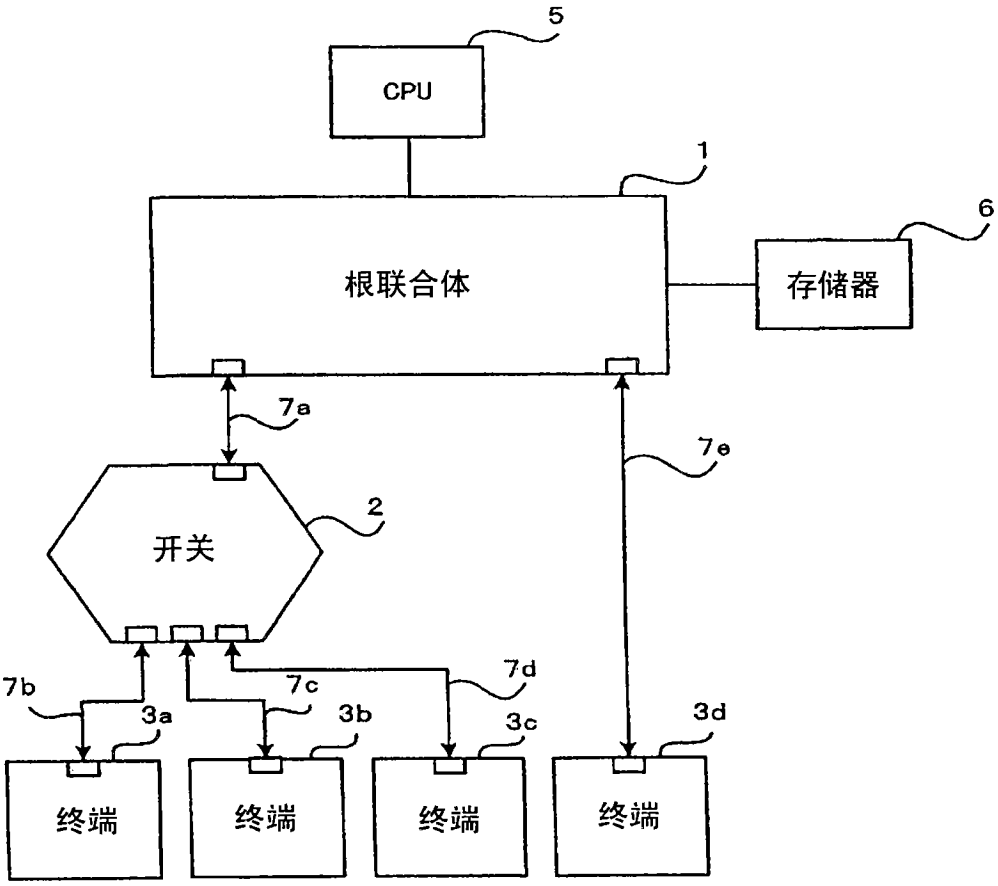

5
CPU
1
根联合体
6
存储器
7a
7e
2
开关
7b
7c
7d
3a
3b
3c
3d
终端
终端
终端
终端

图 1

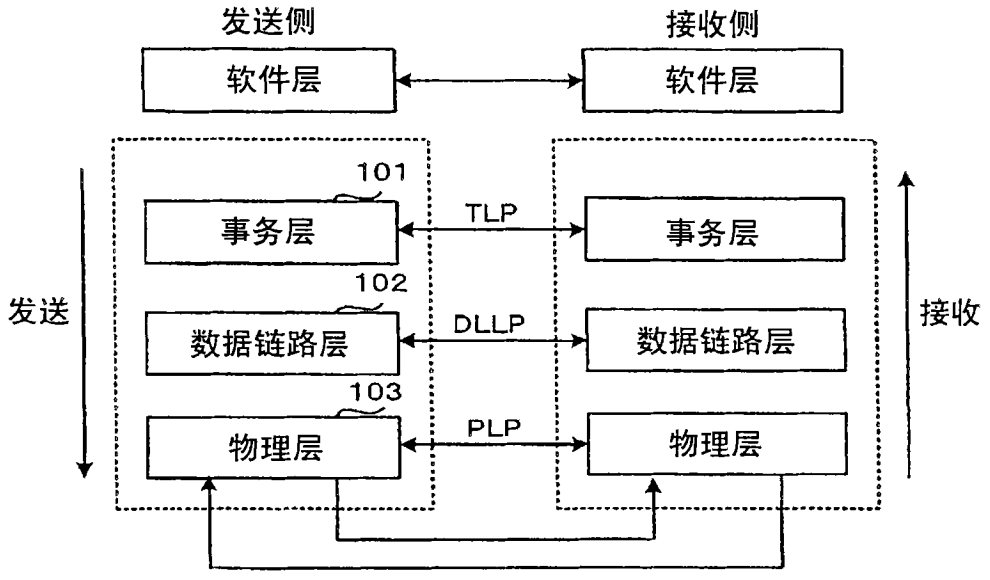

发送侧
接收侧
软件层
软件层
101
事务层
TLP
事务层
102
数据链路层
DLLP
数据链路层
103
物理层
PLP
物理层
发送
接收

图 2

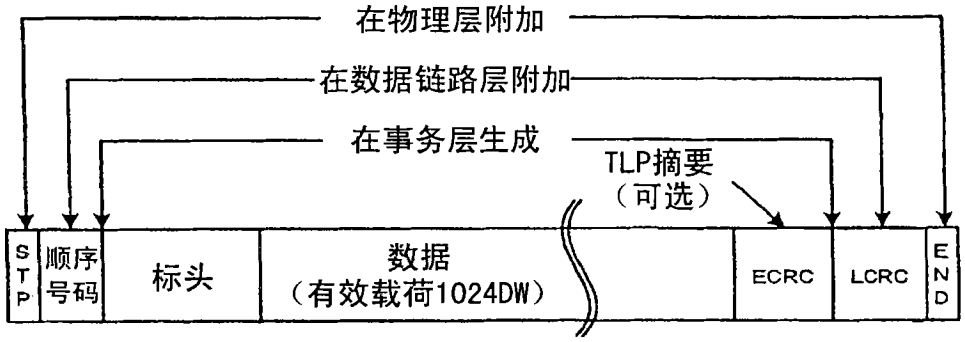


图 3

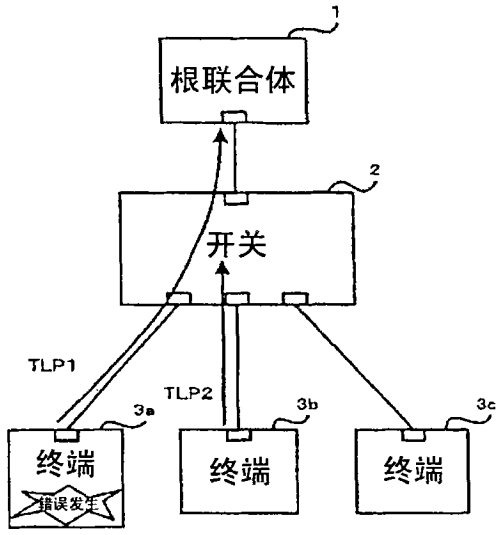


图 4A

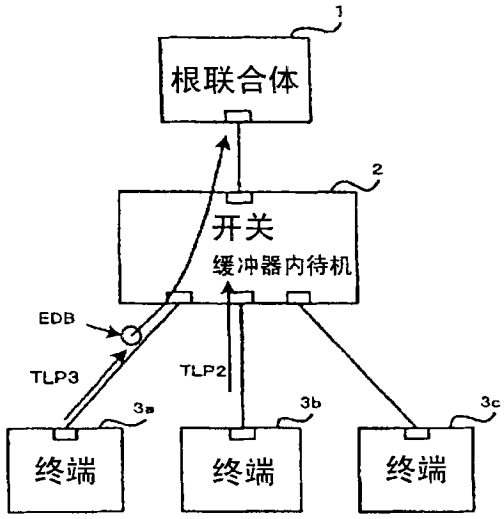


图 4B

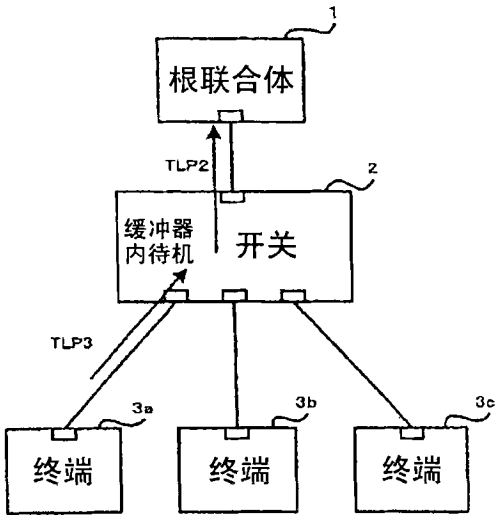


图 4C

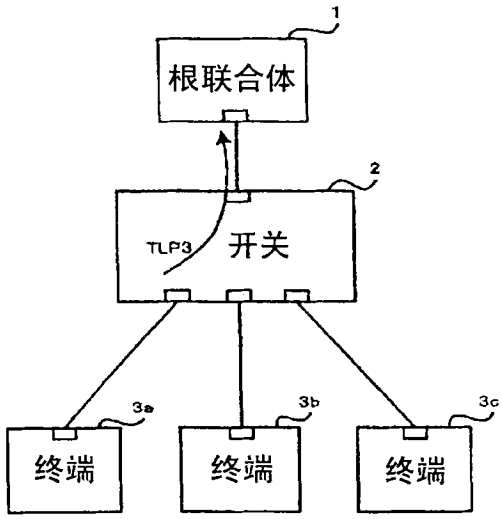


图 4D

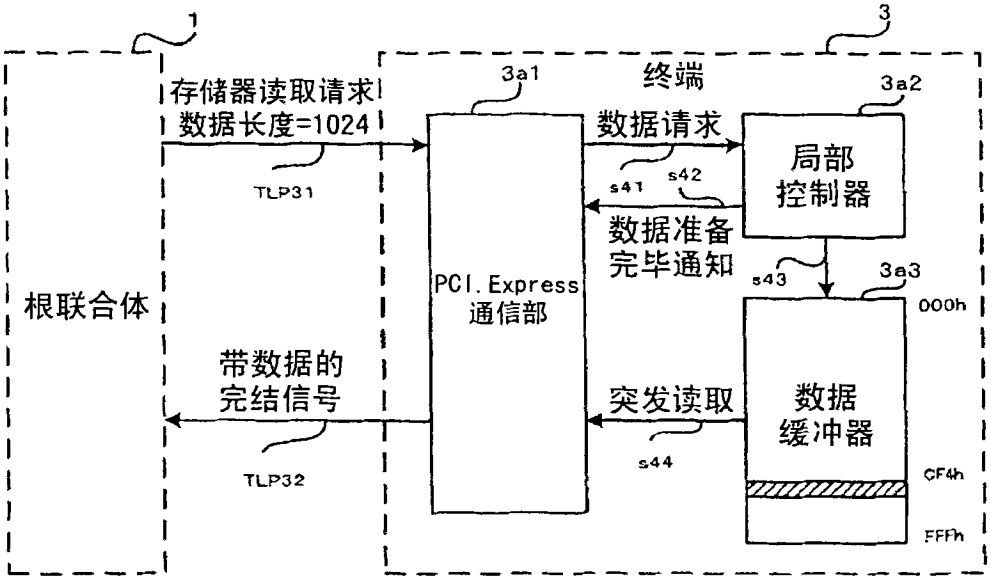

1
3
3a1
终端
3a2
存储器读取请求
数据长度=1024
TLP31
数据请求
s41
s42
局部
控制器
数据准备
完毕通知
s43
3a3
000h
根联合体
PCI. Express
通信部
带数据的
完结信号
TLP32
突发读取
s44
数据
缓冲器
CF4h
FFFh

图 5

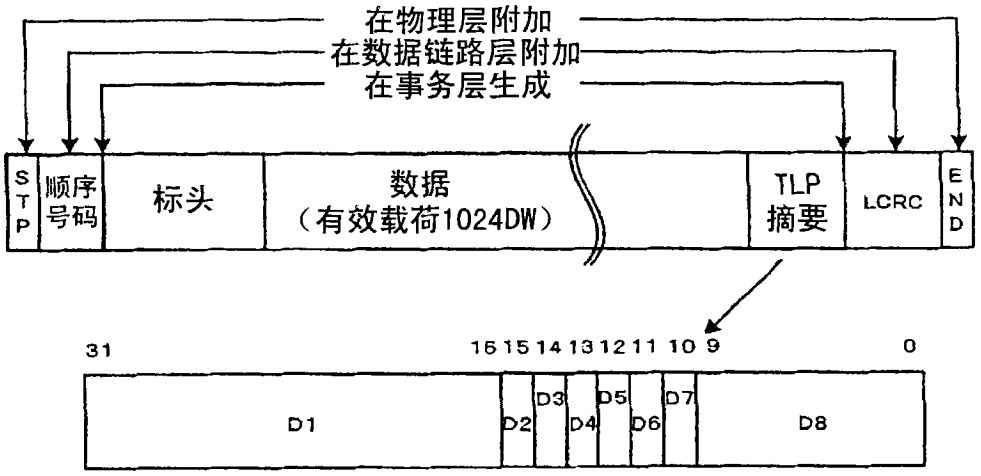

在物理层附加
在数据链路层附加
在事务层生成
STP
顺序号码
标头
数据
（有效载荷1024DW）
TLP
摘要
LCRC
END
31
16 15 14 13 12 11 10 9
0
D1
D2
D3
D4
D5
D6
D7
D8

图 6

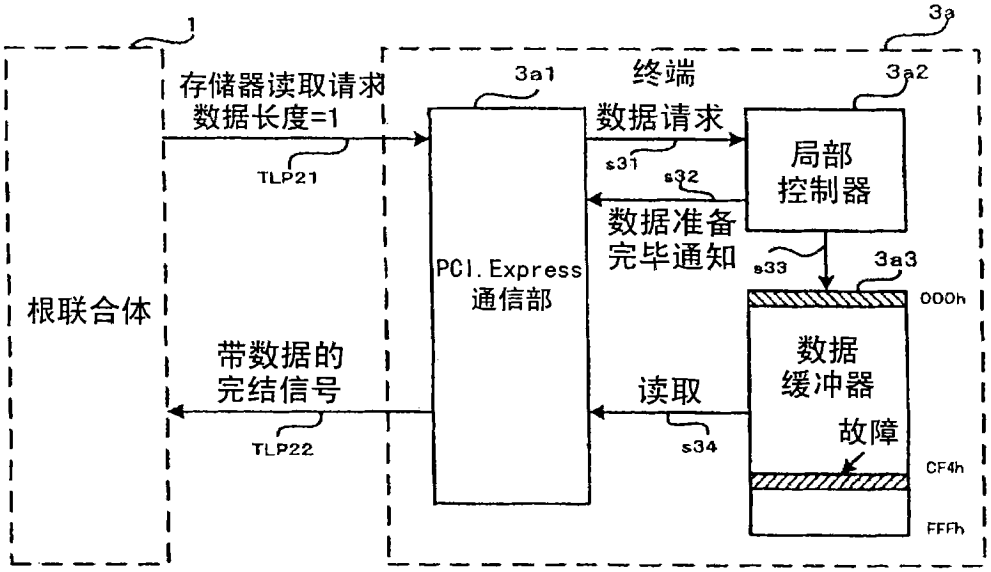

1
3a
3a1
3a2
3a3
终端
存储器读取请求
数据长度=1
TLP21
数据请求
s31
s32
局部
控制器
数据准备
完毕通知
s33
000h
根联合体
PCI. Express
通信部
带数据的
完结信号
TLP22
读取
s34
数据
缓冲器
故障
CF4h
FFFh

图 7

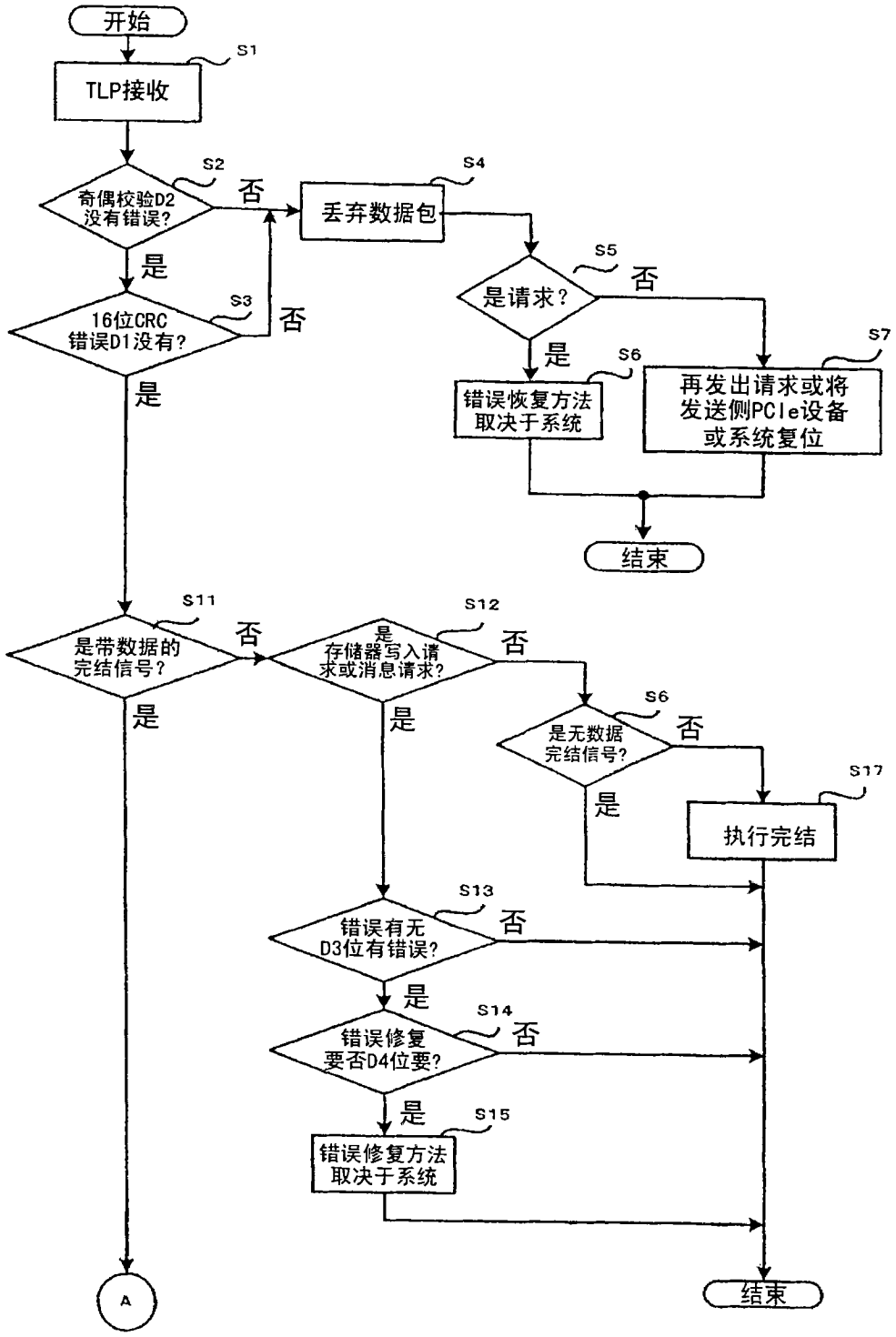

开始
S1
TLP接收
S2
奇偶校验D2
没有错误?
否
S4
丢弃数据包
是
S3
16位CRC
错误D1没有?
否
是
S5
是请求?
否
是
S6
错误恢复方法
取决于系统
S7
再发出请求或将
发送侧PCIe设备
或系统复位
结束
S11
是带数据的
完结信号?
否
是
S12
是
存储器写入请
求或消息请求?
否
是
S6
是无数据
完结信号?
否
是
S17
执行完结
S13
错误有无
D3位有错误?
否
是
S14
错误修复
要否D4位要?
否
是
S15
错误修复方法
取决于系统
A
结束

图 8

图3-A

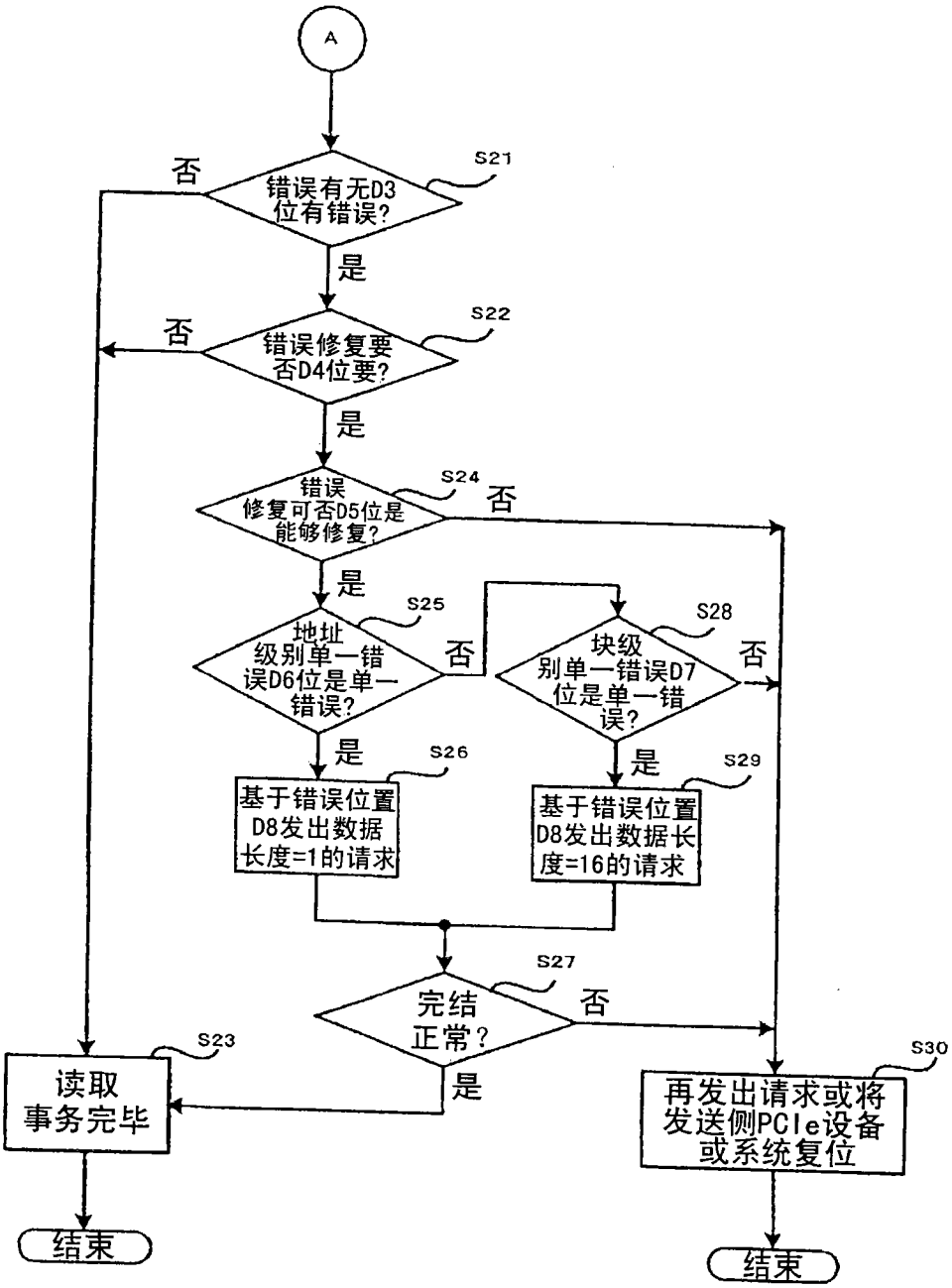

A
S21
错误有无D3位有错误?
否
是
S22
错误修复要否D4位要?
否
是
S24
错误修复可否D5位是能够修复?
否
是
S25
地址级别单一错误D6位是单一错误?
否
是
S26
基于错误位置D8发出数据长度=1的请求
S28
块级别单一错误D7位是单一错误?
否
是
S29
基于错误位置D8发出数据长度=16的请求
S27
完结正常?
否
是
S23
读取事务完毕
结束
S30
再发出请求或将发送侧PCIe设备或系统复位
结束

图 9

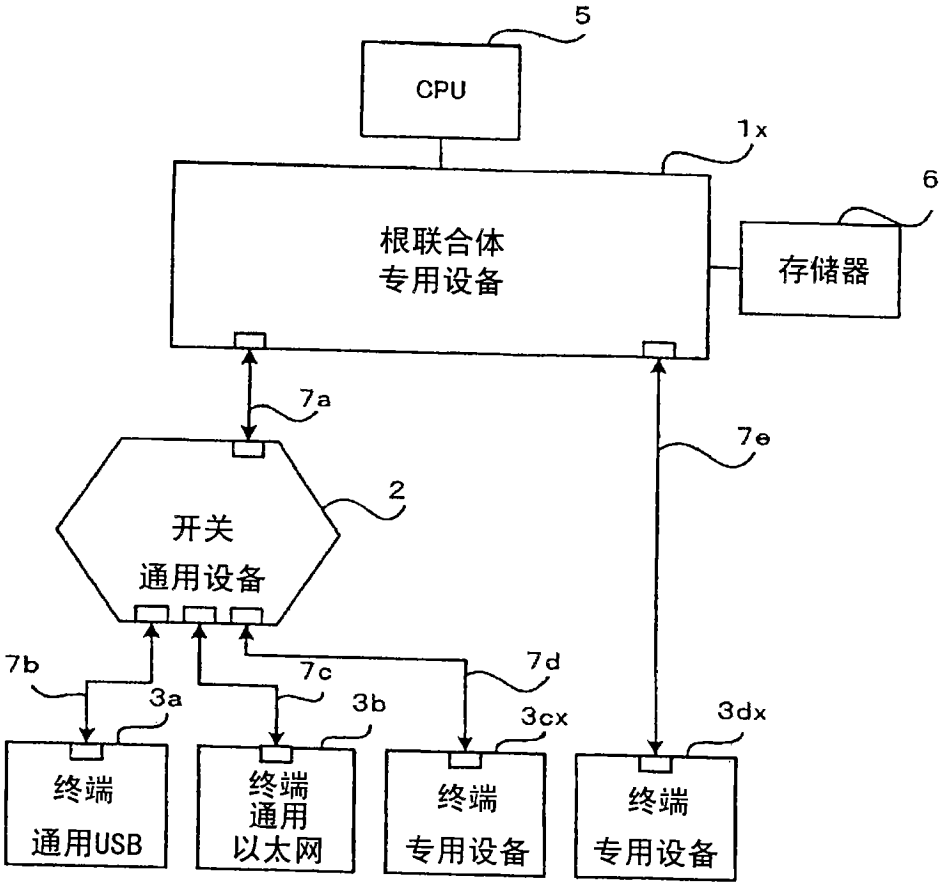

5
CPU
1x
根联合体
专用设备
6
存储器
7a
7e
2
开关
通用设备
7b
7c
7d
3a
3b
3cx
3dx
终端
通用USB
终端
通用
以太网
终端
专用设备
终端
专用设备

图 10